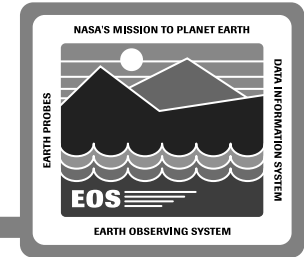


Projected System Access and Utilization

Pitt Thome

System Design Review - 28 June 1994

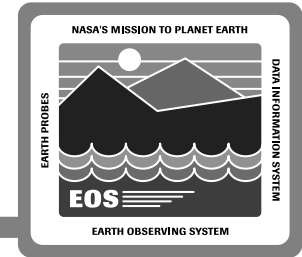
Projected System Access and Utilization



OUTLINE

- Objectives
- Methods
- Assumptions
- Profiles of Projected System Usage
 - How Will Users Enter the System?
 - What Data Will They Use?
 - What Will They Extract from the System?
 - What Are the Inputs to EOSDIS?
- Observations and Implications

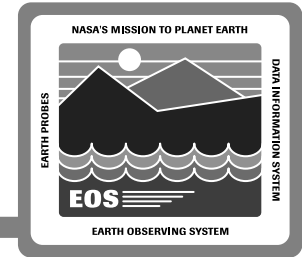
Objective



Characterize projected use of EOSDIS in terms of parameters critical to system design decisions.

- **Bound the problem --- Identify design-drivers to focus future refinements**
- **Provide inputs for engineering analysis:**
 - **software implementation**
 - **hardware topology, sizing and description**

Methods



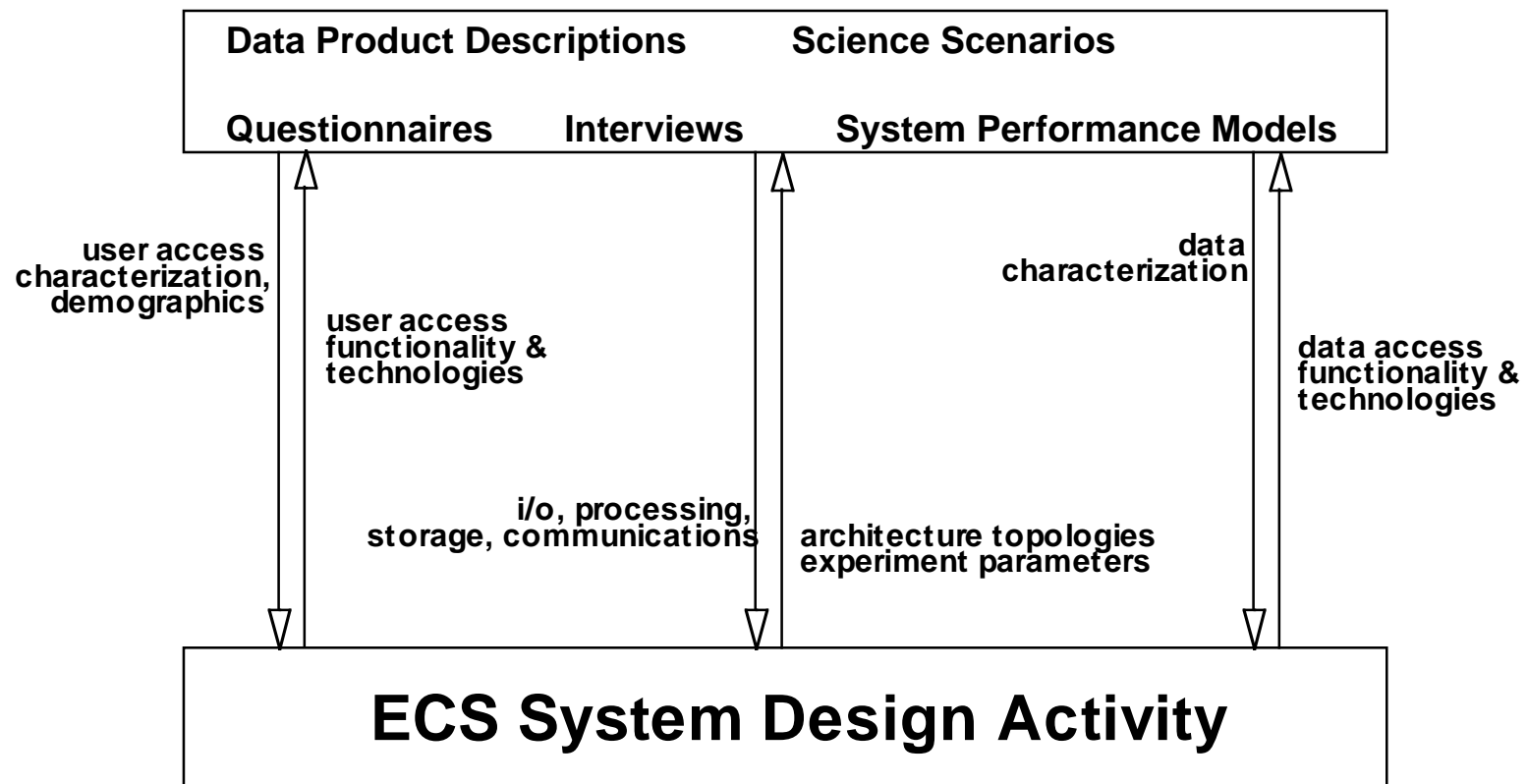
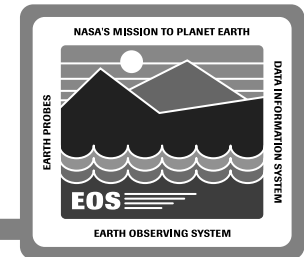
Techniques

- Science Scenario Development
- Questionnaires
- Interviews
- Literature Surveys
- System Analysis
- Static Spreadsheet-based Push, Pull Models
- Quasi-Dynamic C-code Strings Model
- Spreadsheet-based Parametric Access Model

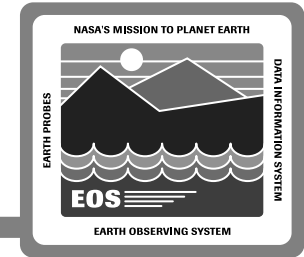
Sources

- Standard Product List
- Market Research Material (i.e., Market Data Retrieval, 1992; Peterson's Guide to Graduate Programs in the Physical Sciences & Mathematics, 1994)
- Existing Data Centers (EDC, NCDC, NGDC, etc.)
- Interviews with Algorithm Development Teams

Inputs to Design



System Access and Utilization Analysis



Surveys-questionnaires-interviews-analyses-modeling-test

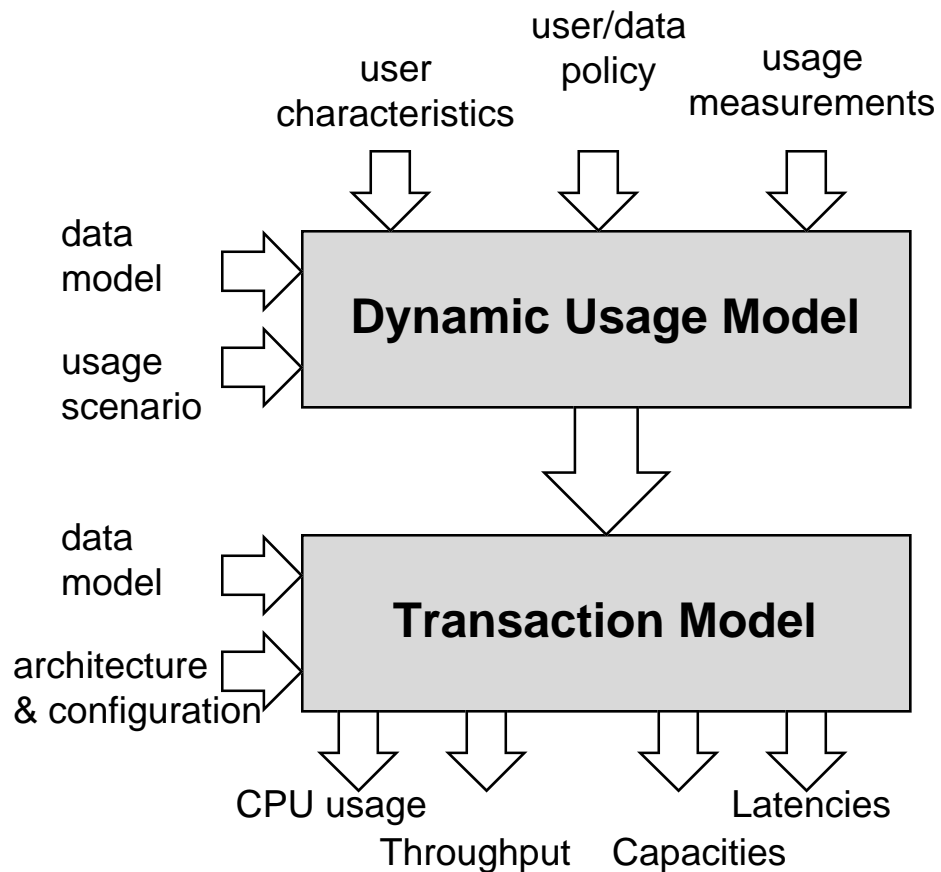
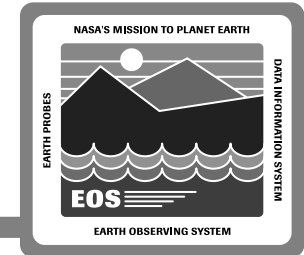
Refine Surveys-questionnaires-interviews-analyses-modeling-test

Refine Surveys-questionnaires-interviews-analyses-modeling test

Next Steps

- **Monitor trends (e.g., Version 0 and prototype usage)**
- **Analyze data collection content, especially browse**
- **Assess effects of various policy options (e.g., pricing) on demand**
- **Refine analysis in areas that are design drivers**
- **Conduct dynamic modeling**

Modeling Plans



• Two Models

- developed in parallel to work together - utilizing lessons learned in early design phase

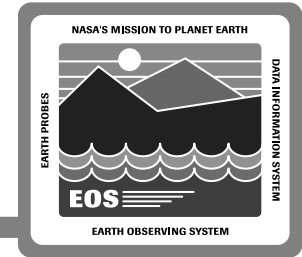
• Usage Model

- includes data/user policy in its characterization
- available at workshops and briefings to investigate usage characteristics
- accounts for requirements from V0 and early releases to assist in future load prediction

• Transaction Model

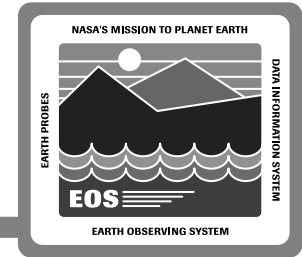
- transaction-based model to determine system loads based on model architecture, and configuration

Assumptions



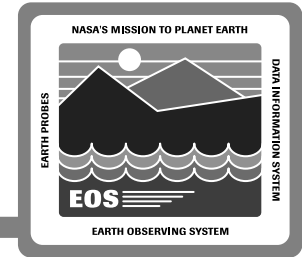
- Current focus is to bound the problem; provide a range of estimates to key questions.
- 1998-2003 - Routine availability of standard products from AM1, TRMM, COLOR, LANDSAT and SAR missions.
- Total cost to EOSDIS users is low enough so as not to influence demand.
- System will provide a suite of services that matches needs of individuals and their communities - would permit management of services as user needs warrant and resources permit.
- Definition of Level 4 products and products derived from scientific investigations to be placed in EOSDIS currently very immature --- expected to significantly increase some demand estimates.

Profiles of System Usage



- **How Will Users Enter the System?**
- **What Data Will They Use?**
- **What Will They Extract from the System?**
- **What Are the Inputs to the System?**

How Will Users Enter EOSDIS?

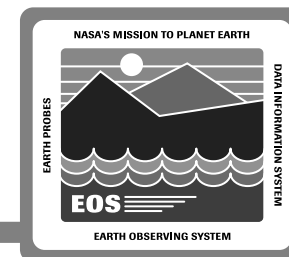


ISSUES

- Potential number of users and frequency of access
- Access methods they will employ
- Access paths to data they will require

IMPORTANCE Permits characterization of system loads and complexity of accesses --- drives sizing of services

Potential Number of Users and Frequency of Accesses



Science Community

Importance: Describes the general landscape and provides overall scope of problem.

Size

United States

NASA EOS Funded Investigators	1900 - 3200
Other Investigators	<u>4200 - 8400</u>
U.S. Total	6100-11600

Other Countries

EOS Investigators	280 - 470
Other Investigators	<u>4000-6000</u>
Other Countries Total	4300-6500

Frequency of Access

Yearly (1-2)	1500	13%
Quarterly (3-11)	3500	30%
Monthly (12-24)	1600	14%
Weekly (25-100)	3400	29%
Daily (100-250)	1600	14%
Total	11600	100%

Total Accesses/year: 545,200

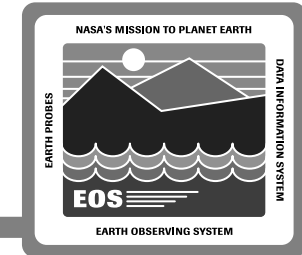
Accesses/Year/User: 47

Implications

Well within predicted technology

Sources: Scenarios, EOS Investigators spreadsheet 4/20/94, 1993 EOS Handbook, Peterson's Guide to Graduate Programs in the Physical Sciences & Mathematics, 1994

Potential Number of Users and Frequency of Accesses



Non-Science Community

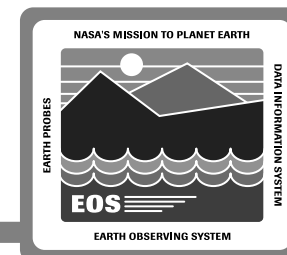
Size	Frequency of Accesses		
<u>United States</u>			
Federal Government	1,500 - 2,200	Yearly (1-2)	186,200 93%
States	1,500 - 3,000	Quarterly (3-11)	10,600 5%
Commercial-End Users,		Monthly (12-24)	800 <1%
Intermediaries, Education		Weekly (25-100)	1400 <1%
Suppliers	400 - 700	Daily (100-250)	1000 <1%
Education (K-12)			
Teachers	2,000 - 7,000		
Students	58,000 - 174,000		
Libraries	6,000 - 12,000	Total Accesses	640,000/year
Policy Makers	TBD		
		Accesses/year/user	3.2
<u>Total</u>	<u>70,000 - 200,000</u>		

Implications

- The potentially large number of Non-Science users calls for:
 - management of resources to allocate ECS services on a priority basis
 - encouraging other service providers to serve these markets
- Design must accommodate interaction with other service providers

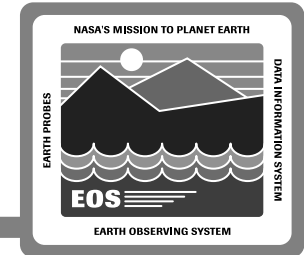
Sources: Questionnaires, Existing Data Centers, Interviews, Market Research Materials

Candidate Approach for Allocating Services

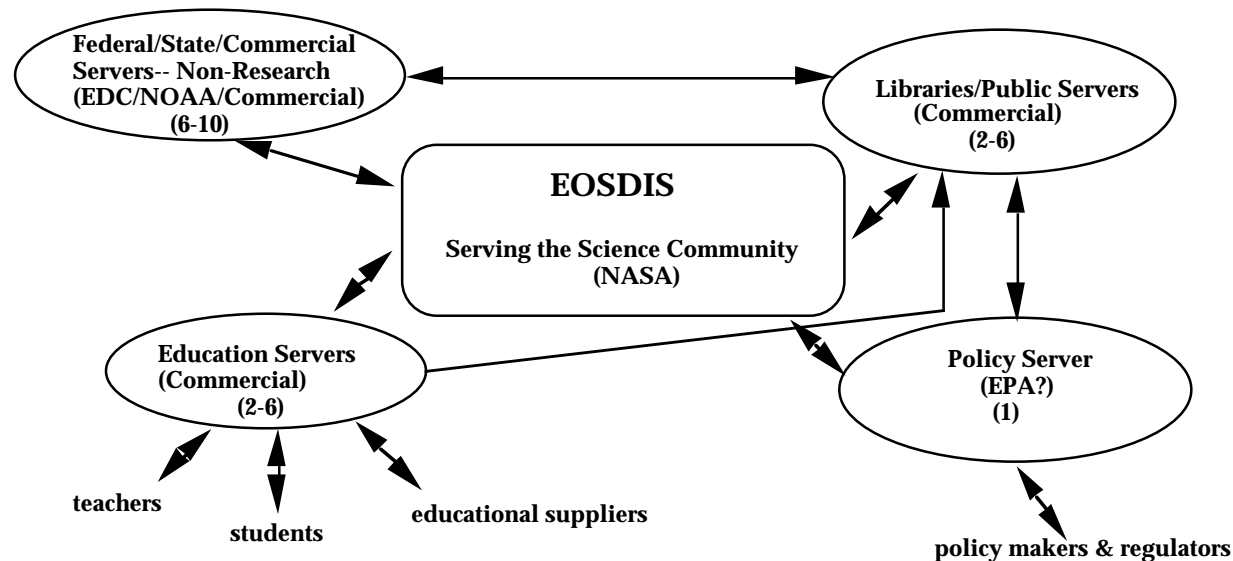


Guest	Registered/External Systems	Registered with Enhancements
Login /authorization	Login /authorization/authentication Profile	Login /authorization/authentication Profile
CHUI/GUI-based I/F to ECS services	CHUI/GUI-based I/F to ECS services	CHUI/GUI-based I/F to ECS services
<u>Search / Browse Information</u>	<u>Search / Browse Information</u>	<u>Search / Browse Information</u>
Service offers	Service offers	Service offers
	Providers	Providers
	Subscriptions	Subscriptions
	Data Dictionary	Data Dictionary
ECS documents	ECS documents	ECS documents
<u>Search Browse Data</u>	<u>Search / Browse Data</u>	<u>Search / Browse Data</u>
Directory	Directory	Directory
Inventory (simple)	Inventory (simple)	Inventory (simple)
	Inventory (complex)	Inventory (complex)
		Inventory via API
Limited FTP retrieval	Unrestricted FTP retrieval	Unrestricted FTP retrieval
	Restricted file access, remote mount, and data server access	Unrestricted file access, remote mount, data server access
Public domain S/W browsers	Public domain S/W browsers	Public domain S/W browsers
	Verify, display, and translate ECS data files	Verify, display, and translate ECS data files
	Basic subsetting and subsampling	Basic subsetting and subsampling
	Interface for user provided	Interface for user provided
	Data analysis and visualization tools	Data analysis and visualization tools
	Feature extraction	Feature extraction
	Enhanced subsetting and subsampling	Enhanced subsetting and subsampling
I/F to electronic mail, BBs, News groups	I/F to electronic mail, BBs, news groups	I/F to electronic mail, BBs, news groups
	Algorithm development tools	Algorithm development tools
	Data acquisition request (DAR)	Data acquisition request (DAR)
		Access to DAAC's
		Data ingest
		Production planning
		Data processing services
		Products under development
Guest Servers	ECS Client Software	ECS Client Software
Tutorial	Tutorial	Tutorial

Service Providers

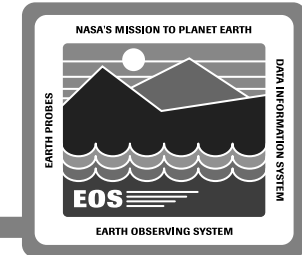


Concept Series of servers providing EOSDIS, value-added, and other products to meet non-science community needs.



Implications Reduces non-science EOSDIS users from 70,000-200,000 to 100-300 users with less demand on resources.

Access Methods They Will Employ



Science

Importance:

- defines what the various loads or accesses are and the loads on the system
- defines what services are needed to support various modes of access
- provides insight into user environments

Entry Through Other Systems

<u>UNITED STATES</u>	<u>NUMBERS</u>
Direct	5,900-11,000
Through other Data Systems:	
NOAA	100-300
CIESIN	?
Other	<u>100-300</u>
Total	6,100-11,600
<u>FROM OTHER COUNTRIES</u>	
Direct	2,900-4,500
Through Other Data Systems (Europe, Japan)	<u>1,400-2,000</u>
Total	4,300-6,500

Access Profile

<u>Method</u>	<u>Percentage</u>	<u>Number</u>
Telephone Interface <u>Only</u>	1.5%	90-170
Electronic	91.5%	5,600-10,600
Subscription		1,400-2,700
Browsers		2,000-3,800
Remote File Access (RFA)		2,800-5,300
Data Suppliers		800-1,500
Machine-to-Machine	7%	430-810
Total	100%	6,100-11,600

(Note: NOT mutually exclusive)

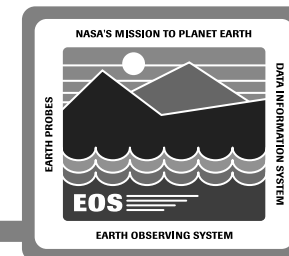
Implications:

The fact that other data centers and individuals from other countries will be accessing EOSDIS indicates that:

- services need to serve heterogeneous communities
- services that allow users to access EOSDIS through other systems need to be provided

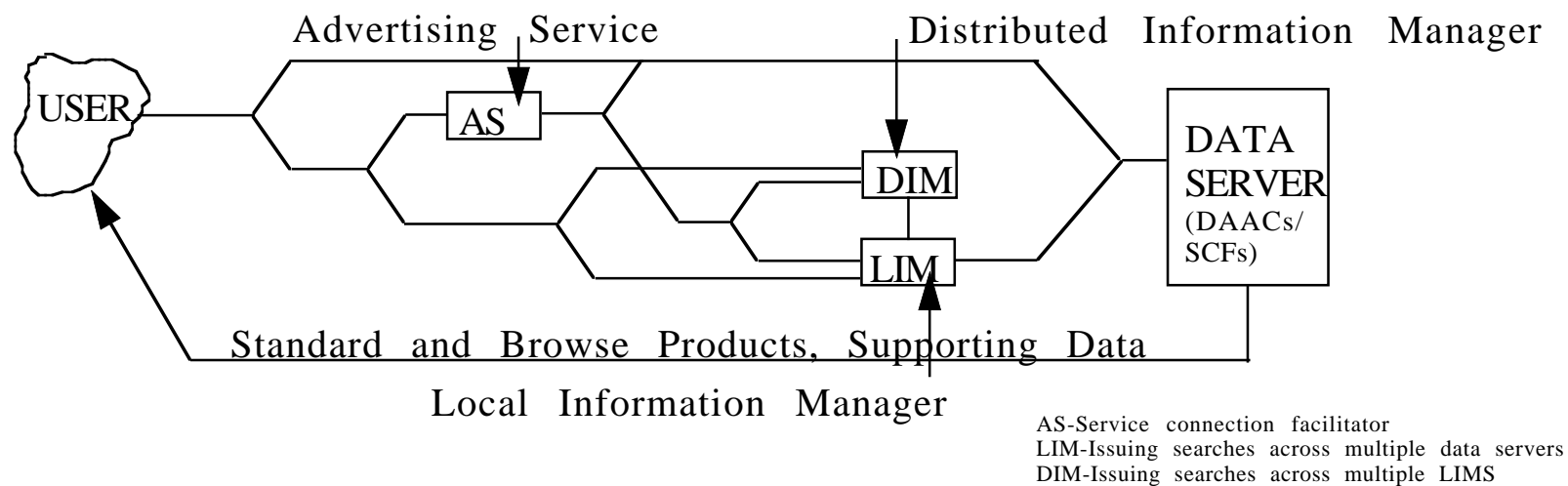
Sources: Scenarios, questionnaires

Access Paths to Data

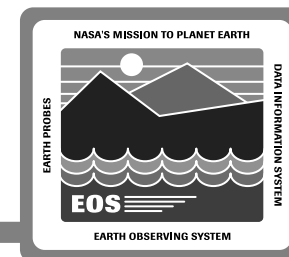


Importance:

Access paths and number of users provide sizing information for various system components



Access Paths to Data



Importance:

Sizing of service components

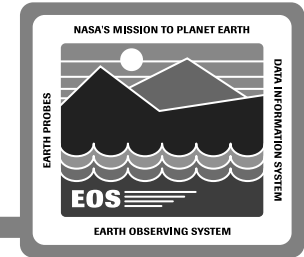
	Accesses/Day	
	<u>Science</u>	<u>Non-Science</u>
Direct to Data Servers	6,000	500
Use of Advertising Service	600	11,200
Through the DIM	100	1,200
Through the LIM	200	1,200

Implications:

- **Attention to services allowing direct use of data servers to support science community**
- **Attention to efficiency of DIM & LIM to support non-science community**

Sources: Scenarios, Questionnaires

What Data Will They Use?



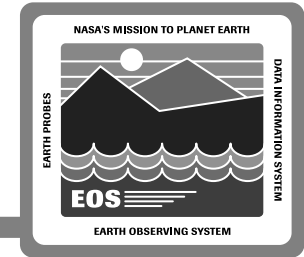
Issues

- Science community's relative discipline focus
- Complexity of search they will employ
- User accesses by Pyramid layer

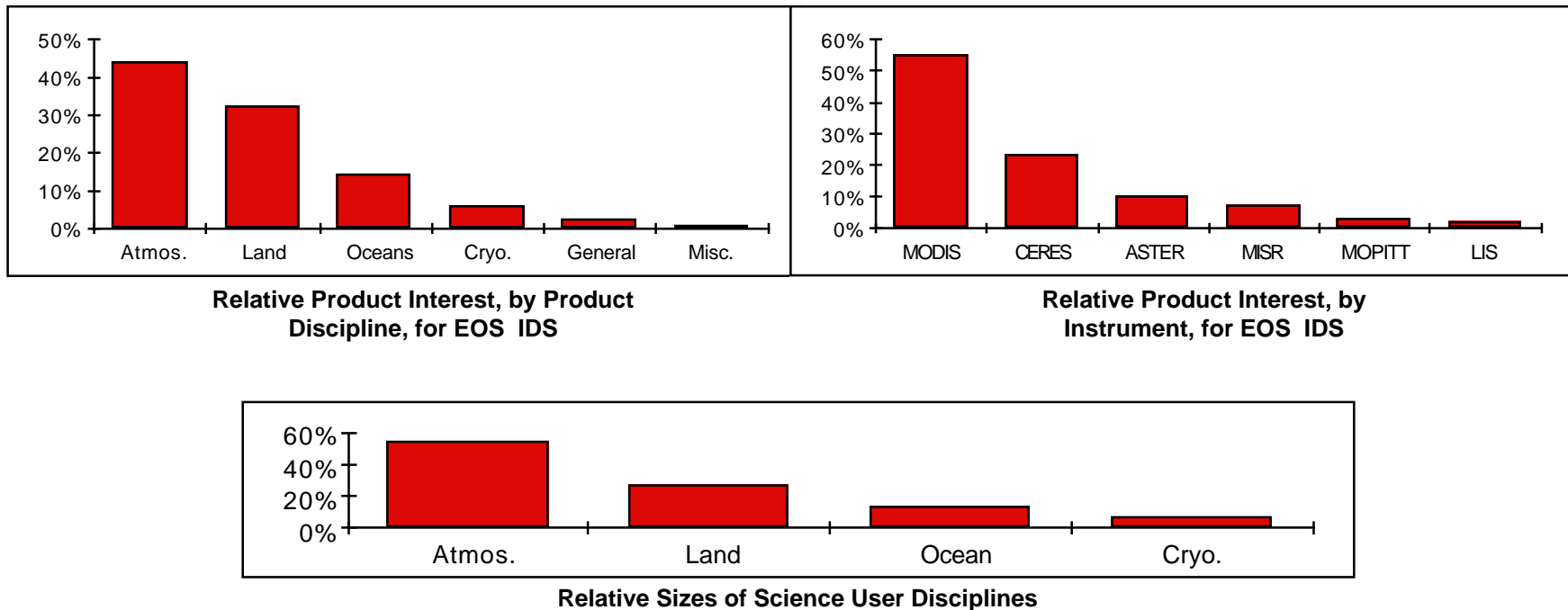
Importance

- Affects disk and tape storage sizing
- Differences in communities' focus imply need for multiple “views” of data
- Accesses of different data types have different performance issues

Science Community's Discipline Focus



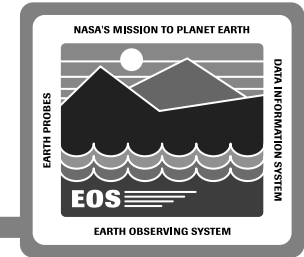
Importance: Understanding the community's discipline focus leads to estimates of relative product demand.



Implications: Several different disciplines will use the same data - ECS must provide customized views of data for different disciplines.

Source: EOS Investigators spreadsheet 4/20/94, Literature Survey, Peterson's Guide to Graduate Programs in the Physical Sciences and Mathematics, 1994

Complexity of Searches



Importance: Understanding complexity of searches allows us to assess loadings on different components of the system (Advertising Service, DIMs, LIMs, Data Server)

	<u>Percentage of Total Searches</u>
<u>Simple Searches:</u>	60-75%
<u>Content Searches:</u>	10-15%
<ul style="list-style-type: none">• Subset• Data Content	
<u>Coincident Searches:</u>	15-25%
<ul style="list-style-type: none">• User Refined Coincident Searches• Match-up Coincident Searches• Complex Coincident Searches	

Note: Not all accesses are Searches.

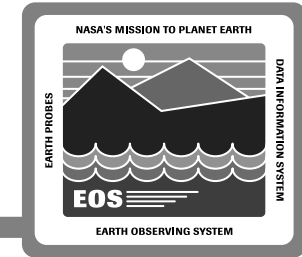
Implications:

- Provide support for different complexity of searches
- Provide efficient support for coincident searches

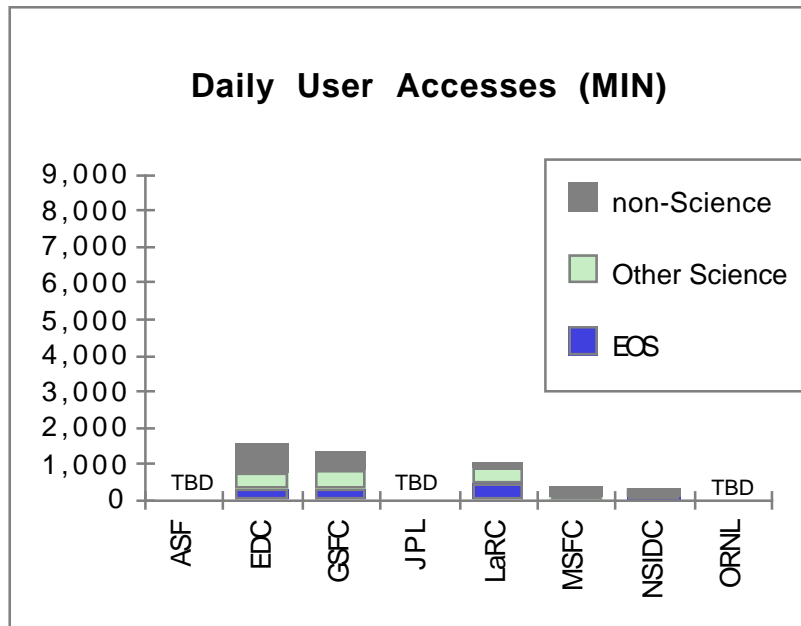
Sources: Scenarios

194-703-PP1-001

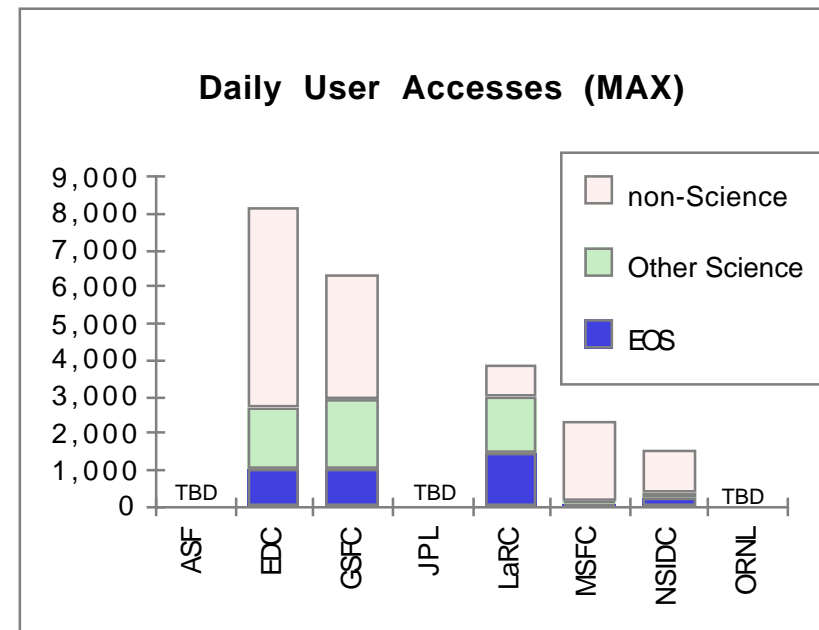
User Accesses to the System by DAAC



Importance: Provides bounding estimates on the level of activity at each DAAC



Total = 4,478 user accesses/day

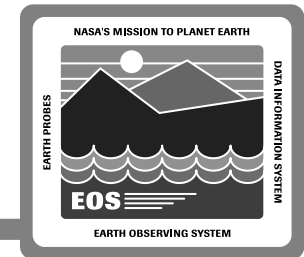


Total = 22,084 user accesses/day

Implications: Heterogeneous access profiles across DAACs: Different design solutions and resource requirements

Source: Scenarios, Questionnaires, EOS Investigators spreadsheet 4/20/94

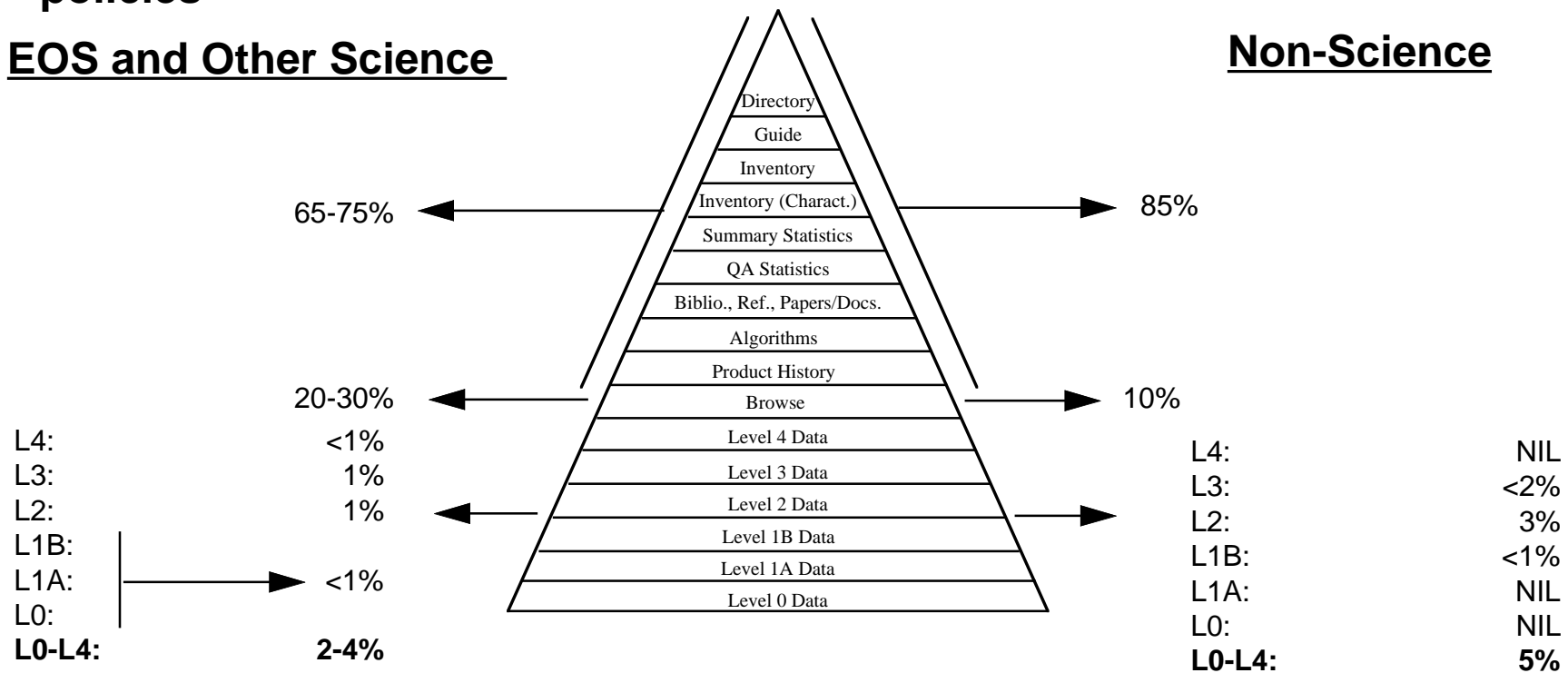
User Accesses to the System By Pyramid Layer



Importance: Selecting the appropriate storage media and staging policies

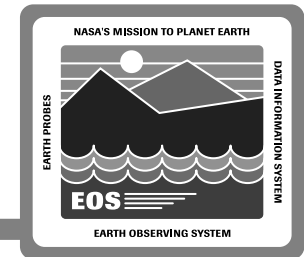
EOS and Other Science

Non-Science



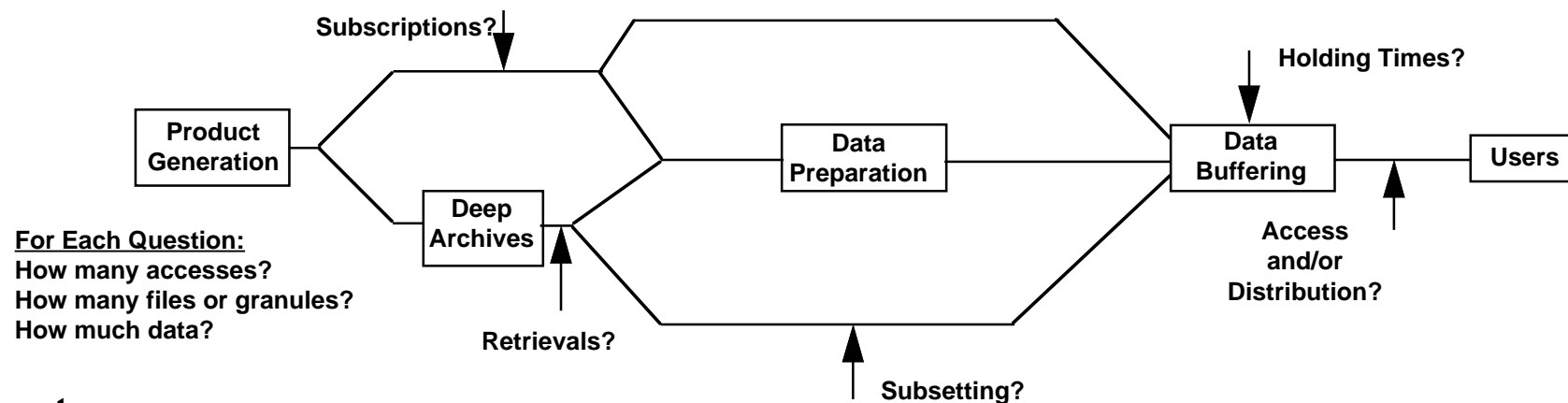
Implications: Confirms the need for rapid access to upper layers

What will they Extract from EOSDIS?



Issues:

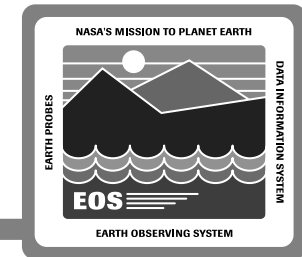
- Volumes by DAAC and Community
- Fraction distributed on physical media and by electronic transfer
- Fraction of distributions via standing order and ad hoc request



Importance:

- Sizing of components
- Lapse time between request and “pickup” required for sizing storage

Geographic Scale of Interest

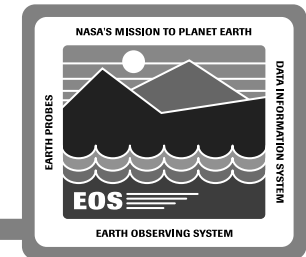


Importance: Volumes are also a function of Geographical Scale

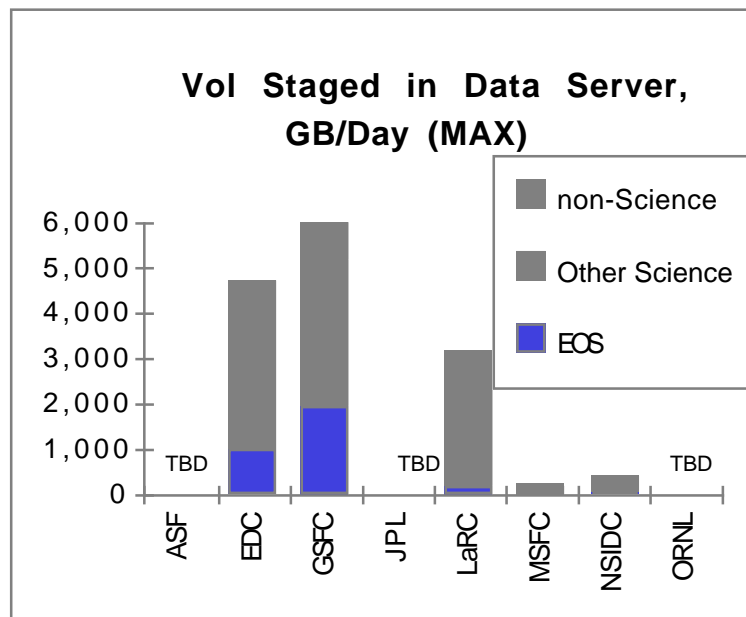
<u>Science Users</u>	<u>PERCENTAGE</u>	<u>NUMBER OF USERS PER YEAR</u>
Browse Products <u>Only</u>	NIL	NIL
1 X 10² KM²	8%	500-900
1 X 10³ KM²	39%	2,400-4,500
5 X 10⁵ KM²	19%	1,200-2,200
1 X 10⁸ KM²	34%	2,000-4,000
Total	100%	6,100 - 11,600
<u>Non-Science Users</u>		
Browse Products <u>Only</u>	93%	14,500-186,800
1 X 10² KM²	3%	2,600-7,200
1 X 10³ KM²	2%	1,800-4,300
5 X 10⁵ KM²	<2%	800-1,300
1 X 10⁸ KM²	<1%	300-400
Total	100%	70,000-200,000

Source: Scenarios, Questionnaires, Interviews
194-703-PP1-001

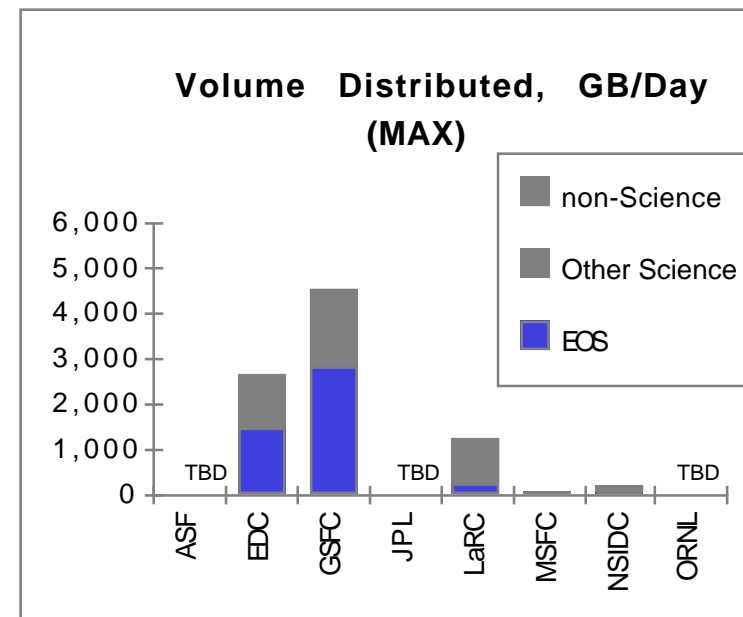
What Will They Extract From EOSDIS?



Importance Impacts storage loading, determines I/O and computing requirements, communications bandwidth



Total = 14,393 GB/day
(5,939 GB/day - min)

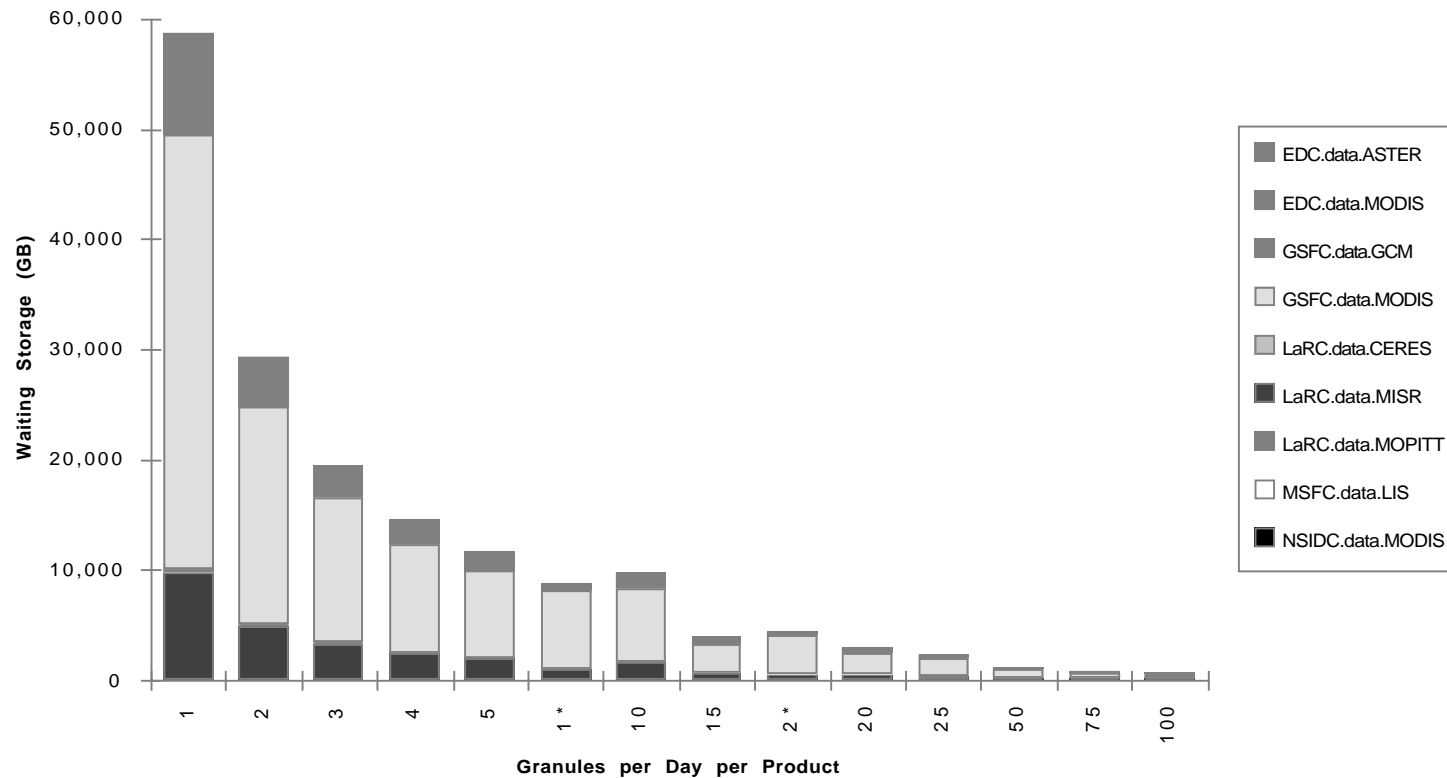
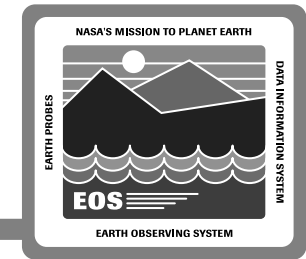


Total = 8,550 GB/day
(3,991 GB/day - min)

Implications

- Volume distributed/volume staged shows amount of subsetting (~50%)
- Each community has different subsetting needs - Variability is from 1/1 to 1000/1

Granule Definition Affects Storage Needs

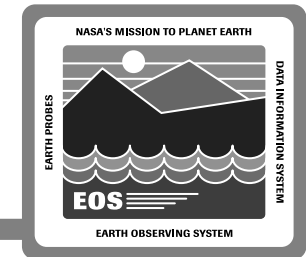


Sources: Quasi-Dynamic C-code Strings Model

194-703-PP1-001

PT-26

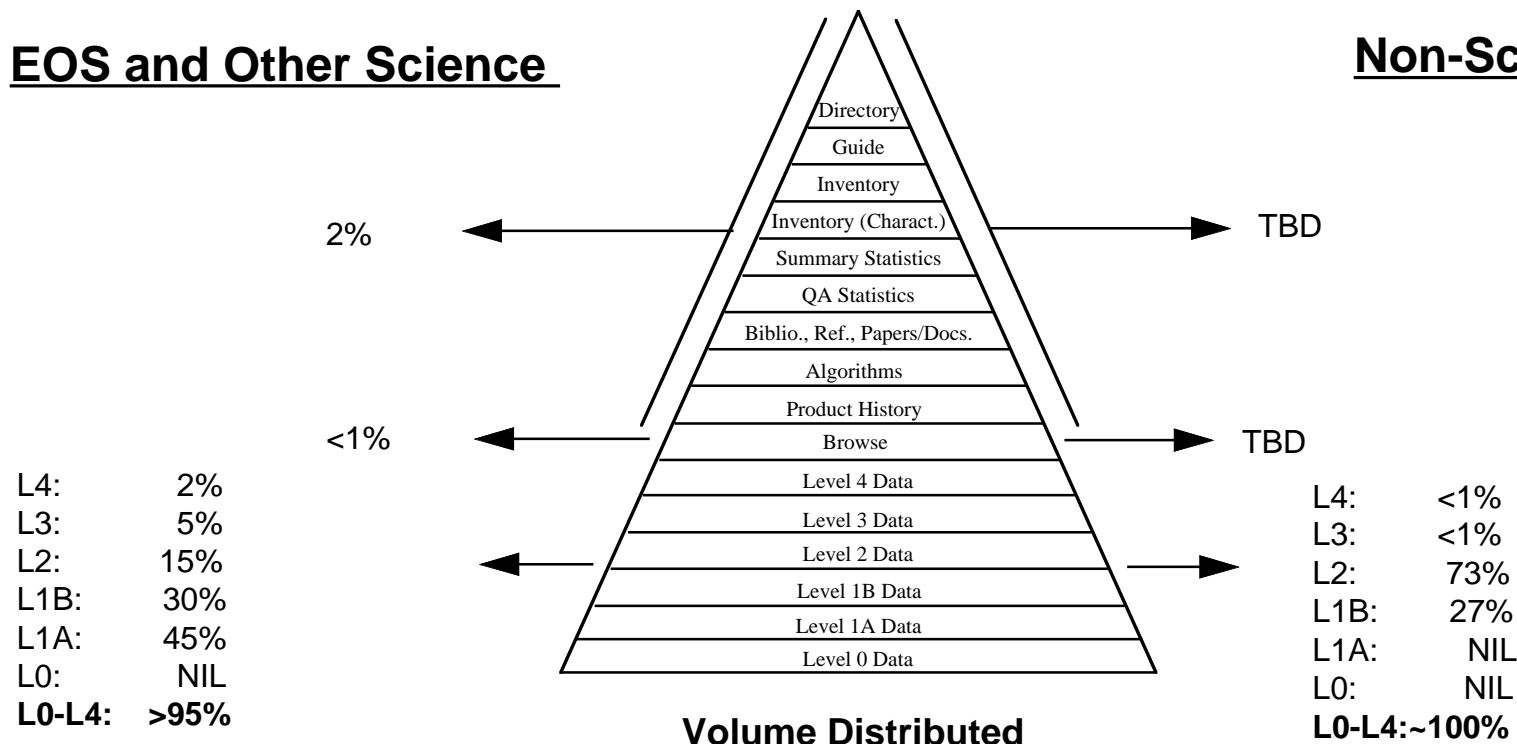
What Will They Extract from EOSDIS?



Importance: Impacts storage, loading, determines I/O and computing requirements, communication bandwidth

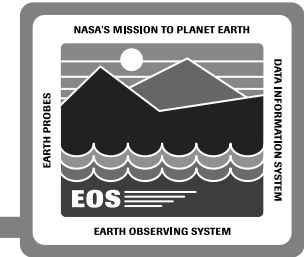
EOS and Other Science

Non-Science



Implications: Confirms need for high I/O bandwidth to access lower levels

What Will They Extract From EOSDIS?



DISTRIBUTION NEEDS

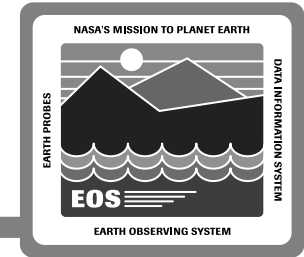
	<u>Percentage of Users</u>
<u>Physical Media</u>	15%
<u>Electronic Transfer</u>	85%
<u>Access Mode</u>	<u>Percentage of Users</u>
• Standing Orders:	30%
• On-Demand/Ad Hoc:	70%

Implications:

- Design must be able to handle high number of electronic distributions, potentially high volumes
- Must look at question of automatic transfer vs. user-initiated transfer
- Significant physical media demand (5%) on standing orders. Therefore, must consider impacts on physical distribution services

Source: Scenarios, Questionnaires

What are the Inputs to EOSDIS?



Current Status of Analysis

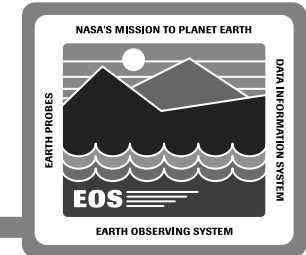
- EOS Platforms
- IDS Level 4 Products and Investigation Results
- TRMM
- Landsat 7
- SAR Missions (E-ERS-2/J-ERS-1/Radarsat)
- Ancillary Data
- Correlative Data
- V0 migration/Pathfinders



TBD

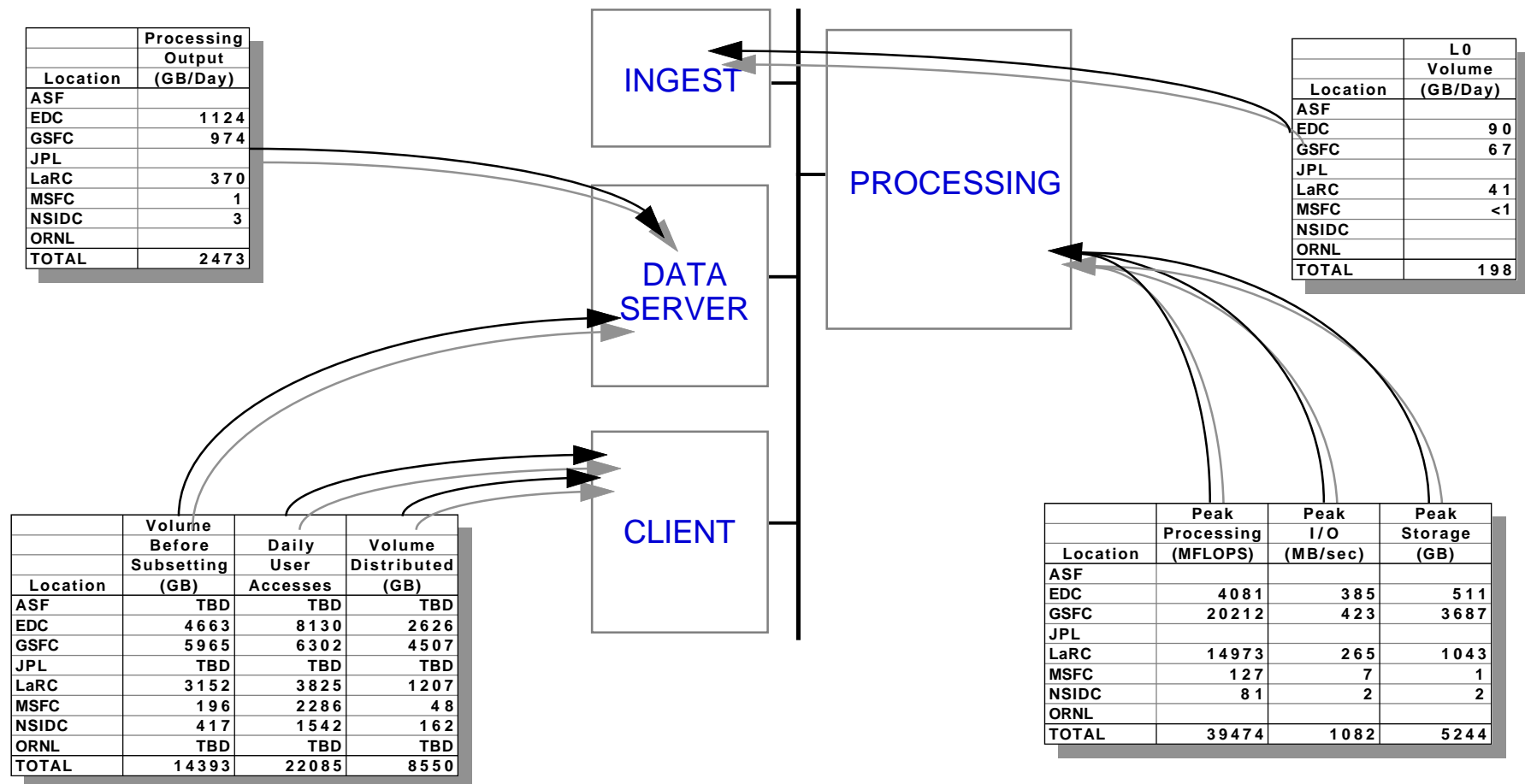
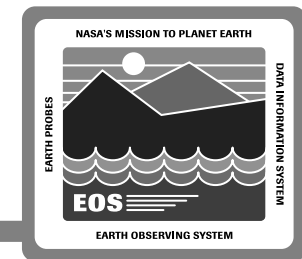


Push Validation Findings

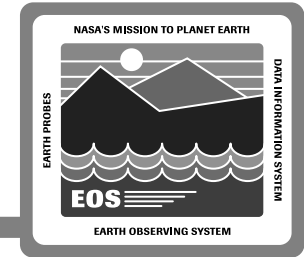


Data Volumes.....	Mostly consistent with SPSO data base, some refinements
Product Definition.....	No Level 3 products at launch
I/O Factors.....	Assumption of one read OK, some data sets read more than once, some only once for more than one product
SCF Interfaces.....	QC reports nominally 1%, problem products, random checks.
Quality Control.....	Most instrument teams are not in line - except MOPITT
Meta Data Generation.....	Estimates included, details of data TBD
Browse Data Generation.....	Estimates included for some instruments, others not
Calibration.....	Done at low frequency (once or twice a month), not a major resource concern (MODIS review TBD)
Use of Quick Look.....	Only ASTER has plans for use of Quick Look in normal operations
Product Granularity.....	Consistent with SPSO data base
Periodic Processing.....	Calibration is only identified periodic processing
On Demand Processing.....	Consistent with SPSO data base, only ASTER

Key Modeling Results



Observations and Implications



Observation Approximately 85% of the users will want distribution of data by electronic means.

IMPLICATION As network capacity grows, ECS must accommodate a large number of users employing electronic transfers.

Observation Almost 30% of the users will request distribution through standing orders.

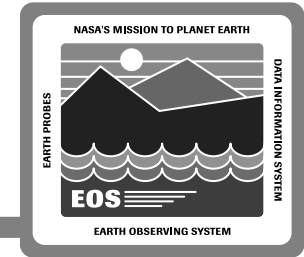
IMPLICATION Provide efficient mechanisms for routine distribution.

Observation Potentially large number of U.S. non-science users (70,000-200,000).

IMPLICATIONS ECS should plan to accommodate interaction with other service providers.

Manage ECS service resources on a priority basis.

Observations and Implications



Observation

The same data is used by a diverse user community.

Diversity of interest extends to scale and nature of queries.

IMPLICATIONS

System must provide views of data customized for different disciplines and communities.

Provide access at the parameter level.

Observation

Large variation in relative dataset interest.

IMPLICATION

Opportunity for improving response time by various techniques for physical placement of data.

Observation

Many users will have difficulty in dealing with information that is very product, discipline, and instrument specific.

IMPLICATION

Design must bridge differences and support system-wide searches.